Grapheme To Phoneme Conversion - An Arabic Dialect Case

Salima Harrat, Karima Meftouh, Mourad Abbas, Kamel Smaïli

GRAPHEME TO PHONEME CONVERSION AN ARABIC DIALECT CASE

S. *Harrat*[1], *K. Meftouh* [2], *M. Abbas*[3], *K. Smaili*[4]

[1]ENS Bouzareah, Algiers, Algeria
[2]Mokhtar University-Annaba, Algeria
[3]CRSTDLA, Algiers, Algeria Badji
[4]Campus Scientifique LORIA , Nancy, France
slmhrrt@gmail.com, Karima.meftouh@univ-annaba.org, m_abbas04@yahoo.fr, smaili@loria.fr

ABSTRACT

We aim to develop a speech translation system between Modern Standard Arabic and Algiers dialect. Such a system must include a Text-to-Speech module which itself must include a grapheme-phoneme converter. Algiers dialect is an Arabic dialect concerned by the most problems of Modern Standard Arabic in NLP area. Furthermore, it could be considered as an under-resourced language because it is a vernacular language for which no substantial corpus exists. In this paper we present a grapheme-to-phoneme converter for this language. We used a rule based approach and a statistical approach, we got an accuracy of 92% VS 85% despite the lack of resource for this language.

Index Terms— *Modern Standard Arabic (MSA), Algiers Dialect,* Grapheme-*to-phoneme conversion, Statistical Machine Translation.*

1. INTRODUCTION

Grapheme-to-Phoneme (G2P) conversion or phonetic transcription is the process which converts a written form of a word to its pronunciation form. G2P conversion is not a simple deal, especially for non-transparent languages like English where a phoneme may be represented by a letter or a group of letters and vice-versa [1].

Unlike English, Arabic is considered as a transparent language; in fact, the relationship between grapheme and phoneme is one to one, but note that this feature is conditioned by the presence of diacritics. The lack of vocalization generates ambiguity at the phonetic level and consequently at the lexcial, syntactic and semantic levels. For exapmle the word "كتب" /ktb/, can have different phonetic transcriptions[1] such as /kataba/, /kutiba/, /kutubun/, /kutubi/, /katbin/... Algiers dialect obeys to the same rule: without diacritics, G2P conversion will be a difficult issue to resolve. That is why we first dedicated our efforts to develop an automatic diacritizer system for this language.

Most of works on G2P conversion have used two approaches: the first one is a dictionary-based approach, where a phonetized dictionary contains for each word of the language its correct pronunciation. The G2P conversion is reduced to a lookup of this dictionary. The second approach is rule-based[2][3][4], in which the conversion is done by applying phonetic rules, which are deduced from phonological and phonetic studies of the considered language or learned on a phonetized corpus using a statistical approach based on significant quantities of data[5],[6].

For Algiers dialect which is a non-resourced language, a dictionary based solution for a G2P converter is not feasible since a phonetized dictionary with a large amount of data is not available. The first intuitive approach (regards to the lack of resource) is a rule-based one, but the specificity of Algiers dialect had led us to a statistical approach in order to consider all features related to this language.

2. ALGIERS DIALECT

The Algiers dialect is one of the most important dialects of Algeria; it represents the dialectal Arabic spoken in Algiers and its periphery. Algiers dialect simplifies the morphological and syntactic rules of the written Arabic. It uses the Arabic alphabet which includes 28 letters (14 solar consonants which assimilate the ل of a preceding definite article ال and 14 lunar consonants which do not assimilate it) and three non Arabic letters ڤ */G/* , ڤ */V/* and پ */P/*. It uses all Arabic diacritics except the Tanween doubled case endings see [7] for more details.

2.1. Writing System

Algiers dialect is a spoken language without conventional writing rules. No writing system is adopted for it by scientific community of the country. In fact, the absence of

[1] We use SAMPA Speech Assessment Methods Phonetic Alphabet for phoeneme representation.

writing system for Algiers dialect is due to the lack of resources such as corpora for this language. Working on this issue (lack of resources) began with a corpus construction which has taken us to adopt a writing system. The main idea for elaborating writing rules is based on the fact that Algiers language is an under-resourced Arabic language, so we adopted the Arabic writing system when it is possible. That means when writing a word in Algiers dialect we look if there is an Arabic word close to this dialect word: if it exists, we adopt the Arabic writing for the dialect word; otherwise, the word is written as it is spoken. It should be noted that Algiers dialect is written from right to left (like Arabic language).

2.2. Issues of G2P Conversion for Algiers Dialect

Algiers dialect G2P conversion obeys to the same rules as MSA. Indeed, Algiers dialect could be considered as a transparent language since alignment between grapheme and phoneme is one to one when the input text is vocalized. This transparency conducts us to adopt a rule-based approach to build a G2P converter. Furthermore, Algiers dialect contains borrowed words from foreign languages (mainly French). Its vocabulary contains many French words used in everyday conversation. French borrowed words could be divided into two categories: the first one includes French words phonologically altered such as the word "فَامِلْيَة" (famille, family) and the second one includes words which are pronounced as in French like the word "سُورْ" (sûr, sure). This last category constitutes a serious deal for G2P conversion since these words do not obey to Arabic pronunciation rules.

Dialect word	Dialect phonetic transcription	French word	Meaning
كوزينَة	/ku:zina/	Cuisine	Kitchen
طابْلَة	/ t'a:bla/	Table	Table
كُونكْسْيُون	/kɔnnɛksjɔ̃/	Connexion	Connection
دُوفِيز	/dəviz/	Devise	Currency

Table 1. Examples of French words used in Algiers dialect

In the examples of Table1, although the first two words are French, they are phonetized as Arabic words. The french word "table" is phonologically altered and written in Algiers dialect (with Arabic script) "طَابْلَة". On the other side, the last two words are phonetized as French words since they are pronounced as in French by Algiers dialect speakers. In order to take account of this word category, the French phonemes like /ɛ/, /ɔ̃/ and /ə/ must be included in Algiers dialect phonemes. Note that, depending on the speaker, we can found different words related to the same meaning falling into both of the two categories. An exemple of this, the word "فَامِلْيَة" (family), which is a french word (famille) phonologically altered is widely used in Algiers dialect, but even the word "فَامِيْ" (famille) with French pronounciation is used.

3. RULE BASED APPRAOCH

As mentioned above the rule-based approach for G2P conversion applied to Algiers dialect requires a diacritized text. Hence, we implemented ADAD, the Automatic Diacritizer of Algerian Dialect texts which then can be converted into their phonetic forms by applying the established rules mentioned in Section 3.2.

3.1. Automatic Diacritic Restoration

ADAD[7] is based on a statistical approach; we used available tools for machine translation where source and target were respectively undiacritized and diacitized texts. We considered diacritization as a Machine Translation problem; we built a Statistical Machine Translation system based on parallel corpora of diacritized and undiacritized texts. We used a trigram language model trained on diacritized texts, a phrase table with undiacritized and diacritized entries. This choice is due to the fact that this approach does not require any linguistic knowledge or resources except vocalized corpora. First, we experimented our solution on MSA diacritized corpora Tashkeela[2] and LDC Arabic Treebank (Part3, V1.0) [8], we got motivating results: a DER (Diacritization Error Rate) of respectively 4.1% and 5.7%, and precision rate of 93.1% and 96%.

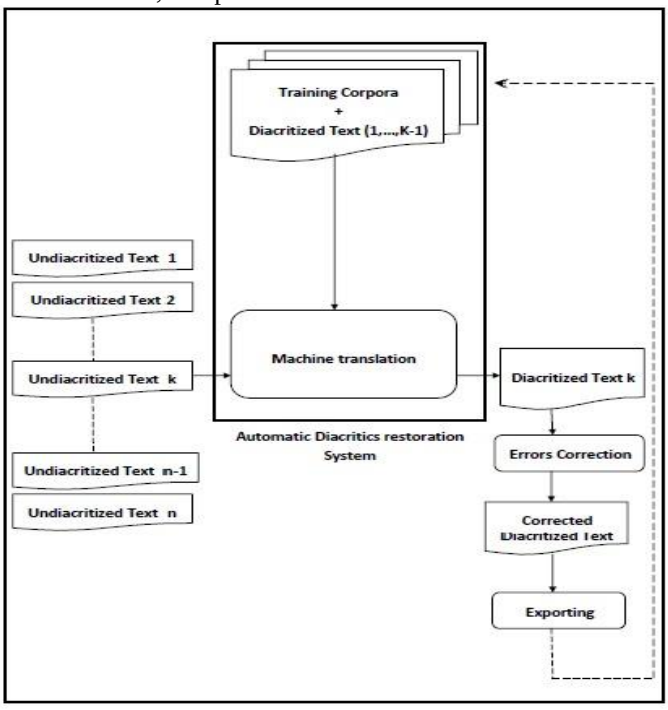

Figure 1. Iterative Diacritization Process

For dialectal side, we vocalized by hand a part of Algiers dialect corpus and we used it for training our system, the results achieved a DER of 12.8% and precision

[2] http://sourceforge.net/project/tashkeela

rate of 98% despite the very little size of this corpus. Considering this interesting precision rate, we used an iterative process to vocalize the rest of the dialectal corpus. We begin by automatic vocalization of a small amount of data, then we proceed to the correction of errors (by hand). This task is not time consuming regarding the high precision rate of the system. The vocalized text obtained constitutes the input to the training corpus then we restart training process, and proceed to automatic vocalization of another portion of the corpus, and so on (see figure 1).

3.2. Rule Based G2P Conversion

After the diacritization step described above, the text is converted into phonemes by applying a set of rules detailed in the following. It should be noted that most of these rules are those adopted also for Arabic [2] [3] and are applicable only for Arabic words and foreign words phonologically altered in our corpus.

Let us consider: BS is a mark of the beginning of a sentence, ES is a mark of the end of a sentence, BL is a blank character, C is a consonant, V is a vowel, LC is a lunar consonant, SC is a solar consonant, and LV is a long vowel. A conversion rule can be written as follows:

LFT + GR + RGT ⇨ /PH/

The rule is read as follows: a grapheme GR having respectively as left and right context LFT and RGT is converted to the phoneme PH. Left and right contexts could be a grapheme, a word separator, the beginning or the end of a sentence or empty.

3.2.1. ذ, ظ and ث rules

In Algiers dialect, the letters ذ /D/, ظ /D‘/ and ث /T/ are not used, they are in most cases pronounced respectively as the graphemes د /d/, ض /d‘/ and ت /t/.

{C, V} + ذ + {C,V} ⇨ /d/
{C, V} + ظ + {C,V} ⇨ /d‘/
{C, V}+ ث + {C,V} ⇨ /T/

3.2.2. Foreign letters rule

Algiers dialect alphabet corresponds to Arabic alphabet extended to three foreign letters g, v and p.

{C, V} + ڤ + {C,V} ⇨ /g/
{C, V} + ب + {C,V} ⇨ /v/
{C, V} + پ + {C,V} ⇨ /p/

3.2.3. Definite article rule

When the definite article "ال" is followed by a lunar consonant (which does not assimilate the "ل") the "ا" is not pronounced.

{ LC}+ال+{BL,BS} ⇨ /l/+/LC/

Example: القْمَرْ (the moon) ⇨ /laqmar/

This rule is the same as in MSA with the difference that in MSA the "ا" is pronounced if the definite article is in the beginning of the sentence.

When the definite article "ال" is followed by a solar consonant the "ل" is not pronounced and the consonant following it is doubled (gemination).

{SC}+ال+{BL,BS} ⇨ /?a/+/SC/ +/SC/
Example : السّقَفْ (the roof) ⇨ /?assqaf/

3.2.4. Words Case-ending

Words case ending in Algiers dialect is the Sukun (Absence of diacritics), so the last consonant of a word should be pronounced without any diacritic.

{BL,ES}+C+{C,V} ⇨ /C/
{BL,ES}+ـْ+C+{C,V} ⇨ /C/
Example: قَبلْ (before) ⇨ /qbal/

3.2.5. Long vowel rules

When ا , و and ي occur in a word preceded respectively by the short vowels ـَ, ـُ and ـِ, their relative long vowels are generated.

{C + ا}+ـَ+{C} ⇨ /a:/
Example : كَاسْ (a cup) ⇨ /ka:s/
{C+و}+ـُ+{C} ⇨ /u:/
Example: فُولْ (beans) ⇨ /fu:l/

{C+ي}+ـِ+{C} ⇨ /i:/
Example: بِيرْ (a well) ⇨ /bi:r/

3.2.6. Glottal stop rule

In Algiers dialect, when a word begins with a Hamza, its phonetic representation begins with a glottal stop. In the end of a word the Hamza is not pronounced.

{C,V}+أ+{BS,BL} ⇨ /?/$
Example: أسْكُتْ (stop talking) ⇨ /?askut/

{BL,ES}+ ء +{C,V} ⇨ /Null/
Example: سماء (sky) ⇨ /sm?/

It should be noted that the hamza in the middle of the word is replaced by the long vowels ا or ي in Algiers dialect. For example the Arabic words "بئر " (a well) and "فأس" (a poleax) correspond respectively to /bi:r/ and /fa:s/.

3.2.7. Alif Maqsura ى rule

Alif Maqsuraa ى (which is always preceded by a fatha /a/) at the end of a word is realized as the short vowel /a/.

{BL,ES}+ى+ {ـَ+C} ⇨ /a/
Example: رْمَى (he threw) ⇨ /rmaa/

3.2.8. Alif Madda آ rule

Alif madda آ is realized as alef /?/ with the long vowel /a:/.

{C}+آ+{C} ⇨ /?a:/

Example: آمَنْ (he trusted) ⇨ /?a:man/

3.2.9. Words ending with ة

The ة is not pronounced in Algiers dialect unlike in MSA where it is realized with the two phonemes /t/ and /h/ (depending on the word position).

{BL,ES}+ة+{C,V} ⇨ /Null/

Example: طِفْلَة (a girl) ⇨ / t'afla/

3.2.10. Words ending with ه

The ه is not pronounced in Algiers dialect when it is preceded by ـُ .

{BL,ES}+ه+{ـُ} ⇨ /Null/

Example: كِتَابُه (his book) ⇨ /kta:bu/

3.2.11. Words containing the sequences ب ،ن

When a ن is followed by a ب , the ن is pronounced as /m/

{ب }+ن+{C,V} ⇨ /m/

Example: مَنْبَرْ (a foretop) ⇨ /mambar/

3.2.12. Gemination rule

When the Shadda ـّـ appears on a consonant, this consonant is doubled (geminated)

{V }+ ـّـ +{C } ⇨ /CC/

Example: سُكَّرْ (sugar) ⇨ /sukkur/

It should be noted that most of these rules could be applied for other Algerian dialects and Arabic dialects close to them such as Tunisian and Moroccan ones.

4. STATISTICAL APPRAOCH

The rule based approach that we adopted in our research didn't take account of French words used in Algiers dialect and pronounced the same way as in French language. Statistical approaches are the best solution to tackle this kind of issues. Again, we considered this situation as a machine translation problem in which the source language is the text (a set of graphemes) and the target language is its phonetic representation (a set of phonemes).

This system uses Moses package [9], Giza++ [10] for alignment and SRILM [11] for language model training.

The main motivation of using a statistical approach is that we can include French phonemes in the training data.

To build this system, the first component is a parallel corpus including a text and its phonetic representation. Actually, this resource is not available, so we created it by using the rule-based converter described above.

This system operates at a grapheme and phoneme level, we split the parallel corpus into individual graphemes and phonemes including a special character as word separator in order to restore the word after conversion process (see example below).

ـْـ	ت	ـُـ	ك	ـْـ	س	ـَـ	ن
Null	/t/	/u/	/k/	Null	/s/	/a/	/n/

5. EXPERIMENTS

5.1. Experimental Material

We used an Algiers dialect corpus constructed by hand. It consists of more than 6K sentences including 10.7k different words. This corpus is a collection of movies and TV shows and some everyday recorded conversations which we transcribed into text (in Arabic script) by hand 8by adopting writing rules. Initially, this corpus was not vocalized, we proceeded to vocalize it as explained in section 3.1.

Our corpus is consisting of three categories of words:

1. Arabic words.
2. French words phonologically altered and their pronunciation is realized with Arabic phonemes.
3. French words for which the pronunciation is realized with French phonemes.

5.2. Results for Rule Based Approach

We applied phonetization rules seen below on a corpus of Algiers dialect words with Arabic words(8993 words) and French words phonologically altered (801 words), the system accuracy was 100%. In addition to Arabic words, French words in the second category are correctly phonetized because their phonetic realization is close to Algiers dialect. For example the word "كُوزِينَة" (kitchen, original French word is "cuisine") which is a borrowed French word phonologically altered is correctly converted as /ku:zina/, while a word in the third category as "كُونِكْسْيُون" (connection, original French word connexion) is incorrectly converted to /ku:niksju:n/ since it is realized /kɔnnɛksjɔ̃/ with French phonemes. Considering these words(906 words), the system accuracy decreases to 92%. The issue of these words is that we cannot introduce rules for French words written in Arabic script, since the relationship between Arabic graphemes and French phonemes is not a one-to-one. For example, the graphemes " ـُو " in a French word written in Arabic script could correspond to the French phonemes /y/, /u/,/ɔ/ or /O/ (see some examples in Table 2).

Dialect word	French phonetic transcription	French word	Meaning
سُور	sʏʀ	Sûre	Sure
پُور	pɔʀ	Port	Port
سُودُور	sʊdœʀ	Soudeur	Welder

Table 2. Examples of mappings between Arabic graphemes "ـُو" and French phonemes

5.3. Results for Statistical approach

In order to evaluate the statistical approach, we needed a parallel corpus (text and its phonetic representation) which was not available. To build this resource we proceeded as follows: we used the rule-based system described below to convert Arabic words and French words phonologically altered (category 1 and 2) into Arabic phonemes. Whereas French words (of category 3) were identified and transliterated into Latin script by hand and then converted to French phonemes by using a free French converter. For example the word "كُونكْسْيُون" is transliterated to "connexion" then converted to /kɔnnɛksjɔ̃/.

For test, we split randomly the parallel corpus into three datasets: training data (80%, 8560 words), tuning data (10%, 1070 words) and testing data (10%, 1070 words). Note that the test set includes 913 words of category 1, 77 words of category 2 and 80 words of category 3.

First we tested the statistical approach on a test set containing only Arabic words and French words phonologically altered (category 1 and 2). We got an accuracy of 93%. Then we proceeded to a test on a corpus including the three words categories, system accuracy decreases to 85%. This result is due to the increase of hypothesis number of each grapheme because of introducing French phonemes in the training data. The graphemes " ـُو " for example in some Arabic words (category 1) are phonetized as the French phonemes /y/ or /ɔ/ instead of the Arabic long vowel /u:/, the phoneme /ɔ̃/ instead of /u:n/. Whereas, some words in category 3 are phonetized with Arabic phonemes by substituting for example the phonemes /y/, /u/,/ɔ/ or /O/ by the /u: /, and /ɛ / by /a:/.

5.4. Discussion

In Table3 we summarize the achieved results. At first glance, and regarding the accuracy rate, we could deduce that rule-based approach is more efficient than statistical approach.

	Statistical approach		**Rule-based approach**	
Test set	Word of cat. 1 and 2	Words of cat.1, 2 and 3	Words of cat. 1 and 2	Words of cat.1, 2 and 3
Accuracy Rate	93%	85%	100%	92%

Table 3. Results Summary

Rule-based approach does not take account of French words of category 3; it achieves efficient results only for Arabic words and French phonologically altered words (category 1 and 2). The results of statistical approach must be analyzed regarding the small amount of training data. Accuracy rate of 85% could be easily improved by increasing the size of training data.

On another side, a hybrid approach could be adopted: instead of using one corpus including all categories of words for training the statistical G2P converter, we can use two corpora: the first one including words of categories 1 and 2, could be processed by rule-based approach. The second corpus is a parallel corpus including words of category 3 with their French phonetization used for training the statistical G2P converter. Unfortunately, we have not sufficient data to test such a converter, since our corpus includes only about 1k words of category 3.

6. CONCLUSION

We presented two approaches to build a G2P converter for Algiers dialect which is an under-resourced language. The first approach is rule based; it gives perfect results for Arabic words and French words phonologically altered. Unfortunately, Algiers dialect includes a big proportion of French words pronounced as in the original language. This category of words could not be phonetized with a rule-based approach since the relationship between Arabic grapheme and French phoneme is not a one-to-one process. We used the statistical approach to overcome this problem. Regarding the small amount of training data, we can consider that the results are acceptable.

To improve the statistical G2P converter, French words pronounced as in the original language must be processed in a separate corpus. That is why we have to develop a module for identifying these words. We can use the rule-based approach for Arabic and French words phonologically altered. This hybrid approach could carry out better results.

In terms of resources, this work allowed us to build a phonetized dictionary for Algiers dialect; at our knowledge no such resource is available at this time.

7. REFERENCES

[1] G. Mininni , A. Manuti A "Applied Psycholinguistics, Positive effects and ethical perspectives". Vol II, 2012.

[2] M. Alghamdi, H. Almuhtasab , M. Alshafi "Arabic Phonological Rules", Journal of King Saud University: Computer Sciences and Information. 16: 1-25. (in Arabic),2004.

[3] Y. A. El-Imam "Phonetization of Arabic: rules and algorithms", Computer Speech Language Volume 18, Issue 4, Pages 339–373, October 2004

[4] M. Zeki, O. O. Khalifa, A. W. Naji, "Development of An Arabic Text-To-Speech System", International Conference on Computer and Communication Engineering (ICCCE 2010), Kuala Lumpur, Malaysia, 2010,

[5] P.Taylor, "Hidden Markov model for grapheme to phoneme conversion", In Proceedings of INTERSPEECH, 1973-1976, 2005.

[6] U. K. Ogbureke, P. Cahill, and J.Carson-Berndsen, "Hidden Markov Models with Context-Sensitive Observations for

Grapheme-to-Phoneme Conversion". INTERSPEECH, page 1105-1108. ISCA, 2010.

[7] S. Harrat, M. Abbas, K. Meftouh, K. Smaili, "Diacritics Restoration for Arabic Dialect Texts", 14th Interspeech, Lyon, France, 2013.

[8] M. Maamouri, A. Bies, T. Buckwalter and H. Jin, "Arabic Treebank: Part 3 v 1.0", Linguistic Data Consortium, Philadelphia, 2004

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. "Moses: Open Source Toolkit for Statistical Machine Translation", Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, 2007.

[10] F. Josef Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51, 2003.

[11] A. Stolcke, "SRILM An Extensible Language Modeling Toolkit on Spoken Language Processing",
Proceedings of the International Conference, volume 2, pp. 901-904, Denver, 2002.